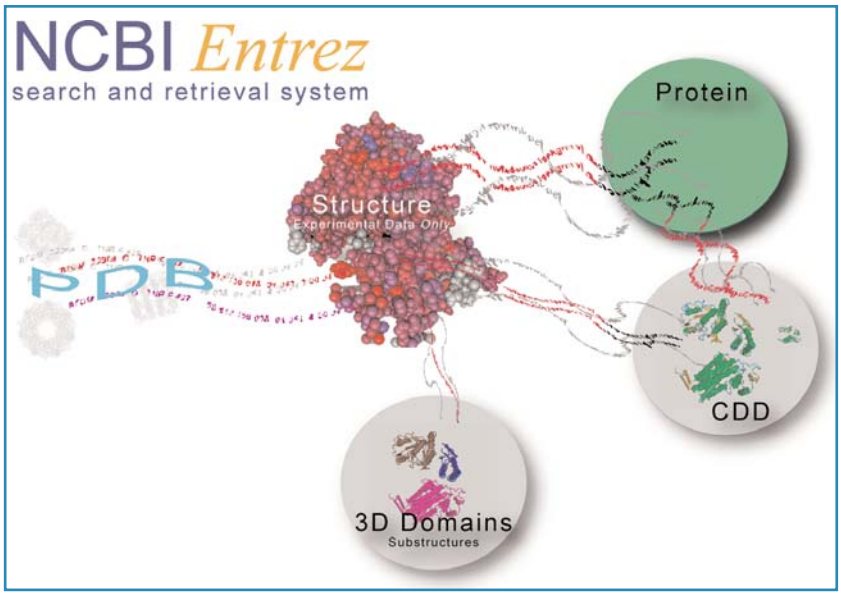
NCBI Structure Resources

National Center for Biotechnology Information ■ National Library of Medicine ■ National Institutes of Health ■ Department of Health and Human Services



www.ncbi.nlm.nih.gov/Entrez/

Entrez Databases and NCBI Tools for Studying Macromolecular Structure

Entrez Databases

Structure - Records contain coordinates, sequence data, and annotations imported from PDB files of experimentally determined structures plus explicit chemical bonding data, uniform secondary structure and domain features, and links to other Entrez databases such as Taxonomy and PubMed. In addition, two coordinate subsets are added to the record: a set containing only backbone atoms, and a set representing a single-conformer model if multiple conformations are present in the PDB file. Protein - Records contain sequence data, feature annotations, and links to other Entrez databases for over 2.6 million proteins from over 61,000 species. Included in this database are all sequences extracted from PDB files.

Domains (CDD) - A collection of protein multiple sequence alignments and any associated structures representing conserved, functional protein domains. Each CDD record is defined by a Position Specific Scoring Matrix (PSSM) that can be used to identify that domain in other protein sequences using RPS-BLAST. Entrez Domain records come from five sources: the Pfam database, the SMART database, the NCBI COG database, the NCBI LOAD database, and the curated NCBI CDD database.

3D Domains - Records consist of compact substructures contained within Entrez Structure records. These domains are identified by searching for breakpoints in the structure between major secondary structure elements so that the ratio of intra- to inter-domain contacts falls above a set threshold. These domains are primarily structural annotations on Entrez Structure records, but can be used in VAST searches to locate similar domains in other structure records.

Tools for Searching Structure Data at NCBI

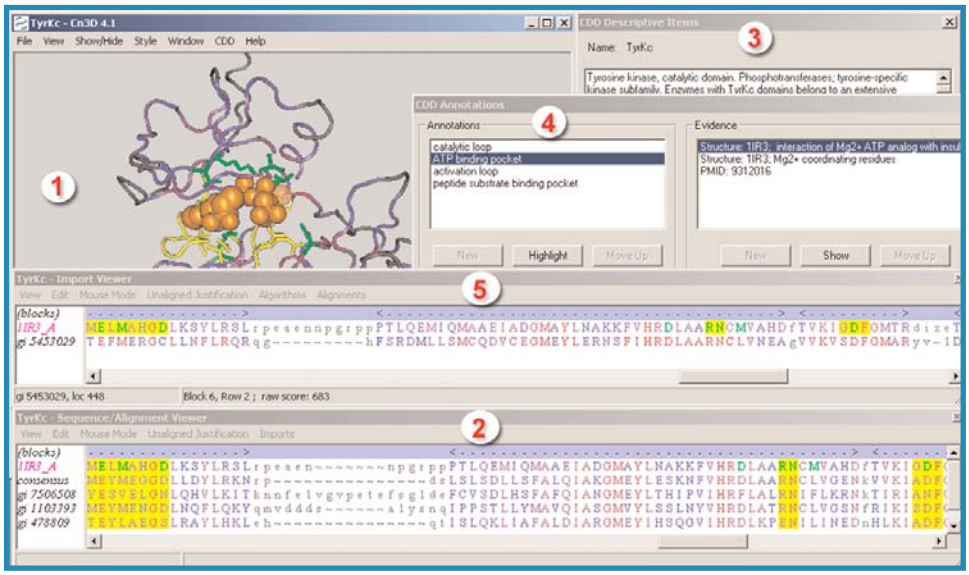
VAST (Vector Alignment Search Tool) - VAST is a tool for finding structural similarity. VAST searches a database of protein structures for records that contain a similar arrangement of secondary structure elements as the query. VAST represents all units of secondary structure as vectors, and then searches for significant alignments among these vector sets. Thus, VAST can tolerate changes in helix or strand length and is insensitive to loop differences, since non-regular structures are not part of the vector set. VAST searches are computed and stored for all Entrez Structure and 3D Domain records, and can also be performed on an input PDB file.
www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml

RPS-BLAST (Reverse Position-Specific BLAST) / CD-Search - RPS-BLAST is a form of BLAST in which a query protein sequence is searched against a database of PSSMs, each representing an Entrez Domains record. This tool thereby identifies conserved domains within the query sequence and provides a statistical assessment of the alignment.
www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

CDART (Conserved Domain Architecture Retrieval Tool) - CDART searches Entrez Protein for sequences that contain the same conserved domains (Entrez Domains) as the query. Those sequences containing the same domains in the same order will be listed first. CDART can also search for specific combinations of domains or for proteins from selected taxonomic nodes.
www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps

Cn3D 4.1: More than just a structure viewer

The links between CDD (the Conserved Domains Database) and Entrez Structure allow users to study and compare the location of functionally important residues in MMDB files. For example, the user might be interested in tyrosine kinases. Let's consider the following query in Entrez Structure: tyrosine kinase site. Among 14 hits, one of the interesting structures is 1IR3 (a human tyrosine kinase from the insulin receptor). The link leads us first to the MMDB page, where the description, references, 3D domains and conserved domains are present, one of which is a curated CD record derived from the TyrKC domain of the SMART database. While the "View 3D structure" button will launch Cn3D showing the structure and sequence of the record, following the link to the curated CD leads to the corresponding CDD page showing the multiple sequence alignment of this domain. From this page, Cn3D can be launched either to show an alignment of multiple structures or to show an annotated view of a particular functional feature in the CD. The figure below displays the ATP binding pocket in this CD.



www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

Five windows are available:

1. Molecular Structure View, showing a 3D model included in the alignment
2. Sequence/Alignment View, showing selected members of the alignment
3. CDD Descriptive Items, providing an overview of the CD
4. Annotations Panel, showing the curated annotations available for highlighting.
5. Import Viewer, showing sequences imported into and then aligned within Cn3D.

Currently the ATP binding pocket has been highlighted, and the indicated residues are colored yellow in the structure and all sequences. The imported sequence (gi 5453029) is a mouse tyrosine kinase aligned to 1IR3 using the BLAST block alignment algorithm. Cn3D has automatically highlighted the position of the ATP binding pocket in the 1IR3 as well as the boundaries of the aligned blocks.

Short description of features present in Cn3D 4.1

Structure Window.

FILE: Open and save structures (ASCII and binary formats); export PNG pictures; recompute structural alignments between multiple structures.
VIEW: Zoom in/out; display and animate multiple structures; spin a structure.
SHOW/HIDE: Select which structures or chains to display, including aligned or highlighted domains or residues; show atoms within a user-defined distance radius from highlighted atoms.
STYLE: Change rendering and coloring styles; create customized annotations of selected regions of a structure.
CDD OVERVIEW: Display the description of the CD, available annotations, and literature references.

Sequence Window.

VIEW: View titles, geometry violations, and self-hit and alignment scores; choose rows to display; find patterns using Prosite syntax; save the current alignment in FASTA, text, or HTML formats; show the taxonomic distribution of the aligned sequences as a tree.
EDIT: Manipulate the aligned blocks using split, merge, create, and delete functions; sort rows by identifier, score, proximity (similarity), or self-hit scores; move all PDB or highlighted rows to the top.
MOUSE MODE: Change the mouse behavior to select rectangles/columns/rows, drag residues horizontally, or move rows.

Import Window.

Many options in this window are similar to the Sequence Window.
EDIT: Import sequences or structures into Cn3D from a local file or over the network.
ALGORITHMS: Choose an algorithm to align the imported sequences to the master sequence: options are simple BLAST, BLAST using the PSSM of the alignment, BLAST block aligner (restricts aligned regions to the current block structure), and sequence-structure threading.
ALIGNMENTS: Merge alignments into the Sequence/Alignment window.

National Center for Biotechnology Information ■ 8600 Rockville Pike, Building 38A, Room 8N-803
Bethesda, MD 20894 Phone: 301-496-2475 ■ Fax: 301-480-9241 ■ e-mail: info@ncbi.nlm.nih.gov
web site: www.ncbi.nlm.nih.gov
02/2003